Unpublished policy trials, the risk of false discoveries and the persistence of authority-based policy

Mendel, Jonathan

Published in:
Evidence and Policy

DOI:
10.1332/174426416X14779397727921

Publication date:
2018

Document Version
Peer reviewed version

"Unpublished policy trials, the risk of false discoveries and the persistence of authority-based policy"

Jonathan Mendel, University of Dundee

Abstract: The use of trials in the policy process is often presented as a move towards evidence-based policy and away from authority-, ideology- or faith-based policy. However, this paper will draw on two UK examples of unpublished policy trials in order to argue that – when policy is based on such research – it should not usually be described as evidence-based. Instead, the paper will argue that policies based on such unpublished trials are, at best, authority-based.

Introduction

There is currently considerable discussion in UK politics – and emerging from the UK Government – about evidence-based policy. There are real hopes that trials and data analysis could enable a more scientific, open and evidence-based approach to policy (Cabinet Office 2012b; Haynes et al. 2012) along with arguments that evidence-based policy requires a move beyond relying primarily on Randomised Controlled Trials (RCTs) (Cartright and Hardie 2012). Data analytics are hoped to offer a better and more 'objective' picture of policies' effects on citizens, transactions and behaviour, while trials are hoped to offer a more effective way of testing and adapting policies (Amoore and de Goede 2008; Hall and Mendel 2012; Haynes et al. 2012). However, this paper will challenge a tendency to simply state that a trial took place and present the results as strong evidence that a real effect was found, which is then viewed as good evidence to base policy upon. Instead, it is important for the details of research design, methodology and analysis to be exposed to public scrutiny so that the quality of the work can be assessed. It is also important to consider the risk of false discoveries in all trials, even impeccably well-designed and conducted projects. This paper draws on examples of two unpublished policy trials in order to illustrate this.

Evidence-based work has long been presented as an alternative to authority-, ideology- or faith-based approaches (for example Gambrill 1999; Marx and Hopper 2005).[i] Advocates of Evidence-Based Medicine have made strong claims about what constitutes good evidence which have implications well beyond the medical field: as Worral (2007) argues, "[r]eal evidence-based medicine results from applying the universal general principles of the logic of evidence to the particular case of medicine". In the context of policing, Sherman (2013: 1) argues that "[i]n contrast to basing decisions on theory, assumptions, tradition, or convention, an evidence-based approach continuously tests hypotheses with empirical research findings". This approach is associated with a move away from authority-based policies which are justified by reference to the expertise or other characteristics of those making recommendations: for example, Davies et al.'s (2000: 3) definition of evidence in the context of evidence-based policy "largely excludes evidence presented in forms such as expert judgement".[ii] For Gerber and Green (2012: 8) experiments are fair in part

insofar as they "involve transparent, reproducible procedures": ideas of openness are a key aspect of work around evidence, trials and policy.

RCTs are often presented as the primary or best a way of doing evidence-based policy: for example, Torgerson and Torgerson (2008: 1) argue that "[t]he 'gold-standard' research methods for addressing the 'what works?' question in 'evidence-informed' policy-making and practice is the…RCT." For Gerber and Green (2012: 7), "[i]n the contentious world of causal claims, randomized experimentation represents an evenhanded method for assessing what works." However, this paper will question some assumptions around this. While there are broader questions about whether randomised controlled trials in general can offer a 'gold standard' for research (summarised in Worral's (2007) discussion of evidence-based medicine) or policy (see Cartright and Hardie 2012) this paper will focus on what happens when unpublished trials are used in policy and on the type of evidence that these trials can provide.

Cartwright (2013) accepts that trials can be a good way of showing that a certain policy works "there", but Cartwright and Hardie (2012: ix) are doubtful of whether policy trials allow a move from knowing that "a policy worked there, where the trial was carried out, in that population" to conclusions about what will work more broadly. This paper, though, will draw on work around evidence-informed policy and evidence-based medicine to consider a preceding question: how reliably some policy trials establish that a certain policy even works 'there'. The paper will use its two cases to argue that trials without adequate, timely publication and scrutiny leave one relying on the authority or eminence of those involved in running the trials; such trials might, at best, allow a form of authority-based rather than evidence-based policy.

Background: positive policy trial findings have a substantial chance of being wrong

Torgerson and Torgerson (2008:1) begin their book on Designing Randomised Trials by noting that a "key reason for undertaking any research is to increase certainty in an uncertain world". However, it is also important to emphasise that an increasing

certainty does not equate to total certainty: even well-designed randomised trials run the risk of being wrong. As Ioannidis (2005) demonstrates, a significant number of positive findings in scientific trials are wrong; he argues that, in scientific research, "most current published research findings are false". For Ioannidis (2005):

> the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05.

More recently, Colquhoun (2014) demonstrates that – where p=0.05 is taken as the threshold for statistical significance – one would expect significantly more than 5% of positive statistically significant trial findings to be false discoveries. While well-designed research might reduce this risk of false discoveries, it will not be eliminated. Statistically significant results from well-designed trials therefore still include a risk of false discoveries, and this can be much higher for smaller and/or lower-quality trials.

However, as discussed below, policy trials may be accepted as robust evidence with minimal public scrutiny and no significant discussion of questions around the risk of false discoveries.[iii] In medicine, there are criticisms of a tendency to adopt a "fetishisation of randomisation as scientific method rather than statistical technique" (May et al. 2005). In the case of some policy trials, the fact a (perhaps problematically) randomised trial has taken place can take priority over broader methodological and statistical issues. If what is presented as evidence-based policy is based on false discoveries from inadequately-scrutinised trials then this is a long way from the intentions of those advocating evidence-based or evidence-informed approaches.

Case studies

This paper springs from previous research failures: I was unable to access details of how trials were used in two UK pieces of policy research.[iv] These two cases are unlikely to be representative of UK policy trials more broadly. However, focussing on two problematic but relatively high-impact cases is sufficient for the paper to demonstrate that particular risks can be associated with such trials.

This article will consider Behavioural Insights Team research where, despite large reported effects, there was not a timely publication of the detail of the research's methodology and analysis; it therefore could not be publicly assessed when results were announced (see Cabinet Office 2012a). The article will also discuss a Her Majesty's Revenue and Customs (HMRC) trial on using credit referencing data to challenge fraud and error in the tax credit system (see Ross 2011). In the case of this trial, the information which has been released raises serious questions about the quality of this research. In both cases, the Government's refusal to share information about the trials' methodology and analysis in a timely way made it impossible for external scholars to critically assess or replicate the studies, and therefore to confirm or refute their validity.[v] While I attempted to make this information public via Freedom of Information (FOI) requests, this was unsuccessful. The known flaws of the HMRC trial and opacity of both cases raise questions about government trials and, more generally, about evidence-based policy initiatives and the scientific-sounding claims on which they rely.

Behavioural Insights Team: increasing tax repayments

The Behavioural Insights Team, previously part of the Cabinet Office, has worked on how RCTs might guide public policy.[vi] When researching how to get people to pay tax more promptly they reported finding ways to increase tax repayments by 15% (Cabinet Office 2012a). This sprung from three trials where, in the first RCT, "a range of different messages [some incorporating social norms] were tested in letters sent to 140,000 taxpayers...there was a 15 percentage point increase from the old-style control letter which contained no social norm and the localised social norm letters". A second RCT contacted 1,400 late payers with letters that "sought to test the effectiveness of a localised norms message in three larger localities" and a "third RCT

was conducted on 108,000 self assessment debts, worth £290 million. This trial tested a control letter against five other letters to examine the effectiveness of messages highlighting the impact of paying tax on public services in terms of gains or losses, as well as generic and 'injunctive' norms" (Cabinet Office 2014: 22-4). The first RCT reported a 15% increase in responses "from the old-style control letter which contained no social norm and the localised social norm letters" while preliminary results from the second trial are reported as showing that "payment rates rose from 38.7% to 45.5%" (Ibid). For the third RCT, the Cabinet Office reported that "[p]rovisional results suggest that a social norm message is much more effective if it contrasts the recipient's behaviour with the norm" (Ibid).

The Behavioural Insights Team argue strongly for publication of policy trials: "[w]hen any RCT of a policy is completed it is good practice to publish the findings, with full information about the methods of the trial so that others can assess whether it was a 'fair test' of the intervention. It is also important to include a full description of the intervention and the participants…Following the CONSORT statement will ensure the key parts of the trial and the interventions are sufficiently accurately described" (Haynes et al. 2012: 31; see also Schulz et al. 2010). However, the Behavioural Insights Team's initial reporting of these aforementioned three trials was very brief: excluding one box and two figures, they are covered in 701 words (Cabinet Office 2012a: 22-4). There is limited detail about methodology: for example, how sampling was carried out or the precise changes made to test conditions. There is also a lack of detail on the types of analyses carried out. These write-ups neither meet nor aim to meet CONSORT's influential standards for reporting. For example, they do not meet CONSORT's call for details of: "Statistical methods used to compare groups for primary…outcome(s)…Methods for additional analyses, such as subgroup analyses and adjusted analyses" (Schulz et al. 2010: 699). A FOI request intended to obtain more detail about this research published was refused, with the Cabinet Office (at internal review) referring to planned future publications, including publications in academic journals, as a justification for the delay (What Do They Know 2013).

In 2014 (more than 2 years after the Behavioural Insights Team announced the results discussed above) Hallsworth et al. (2014) published a National Bureau of Economic Research (NBER) Working Paper.[vii] The field experiments that Hallsworth et al.

(2014) outline seem to include at least part of the research that the Cabinet Office (2012a: 22-4) has reported but (despite contacting the Behavioural Insights Team) I was unable, at that time, to get clarification of how this related to the original research.

The Behavioural Insights Team made significant claims for what some of its research found, without publishing the information needed to critically assess these claims in a timely fashion. This is despite a Behavioural Insights Team publication arguing that publishing trial protocols *before* the trials start, "so that people can offer criticisms or improvements before the trial is running", is desirable (Haynes et al. 2012: 31). In this instance, though, the policy of methodology being available in advance was not achieved in practice (though the Behavioural Insights Team are preparing to publish full methods in journal articles[viii]) (Hallsworth 2016).[ix] The fact delays in journal publication are (as Hallsworth (2016) notes) common and can be rather long means that, particularly once results have been publicised, it is problematic to withhold information based on such planned publications.

When the opportunity for timely public review and replication of work is removed without good reason, due to delayed publication, it is not appropriate to class policy based on this work as evidence-based.[x] While there is internal scrutiny of government research, publication allows important additional scrutiny and input (see Haynes et al. 2012: 31). Without publication, the RCT evidence in question only has value insofar as one trusts the researchers and internal scrutiny processes involved, based on little information about how the research has been conducted. However skilled the researchers and internal reviewers are, no research is perfect and even the best researchers make mistakes. We are therefore left with something much closer to an old-fashioned reliance on expertise or eminence: one relies on the authority of the researchers and others involved in the project. It would therefore be better to call policy based on such work authority-based (due to this reliance on authority) rather than evidence-based.

HMRC: credit referencing data, tax credits, fraud and error

HMRC has sought to draw on Experian credit referencing data in order to deal with – and reduce losses from – fraudulent and erroneous claims for tax credits. A FOI response from HMRC states that they began their trial with a sample of 20,000 cases "identified using internal HMRC 'risk rules' that suggested there may be another adult living at the same address as the claimant." Subsequently, a group of "600…cases [that] were taken from those graded as 'high risk'…and 150 from those graded as 'medium risk'" was selected. The use of credit referencing data was trialled on this group.

This use of credit referencing data was supported in a *Telegraph* article by the argument that "figures showed that £5.2 billion was lost to fraudulent benefit claims and administrative errors…A trial has already saved £16 million after payments often exceeding £40,000 were stopped" (Ross 2011). Government agencies thus hope that actions can be driven by what can be inferred about people by drawing on relevant data (see Amoore 2011). The same *Telegraph* article confidently offers quantitative predictions of future savings: "The contract is expected to save [HMRC], which administers the tax credits system, £700 million, while the Department for Work and Pensions aims to avoid £100 million in potential losses through the overpayments" (Ross 2011). Such quantitative claims are attractive: as Porter (1995: 74) observes, "Strict quantification, through measurement, counting and calculation, is one of the most credible categories for rendering nature or society objective."

This trial has not been published so FOI requests were used to try to access information regarding this trial. I had limited success, and the information which has been made available raises serious questions about the trial's quality.

Withheld information

The public reporting of this trial – in a fairly brief *Telegraph* article – offers little detail: the article simply stated that a "trial has already saved £16 million after payments often exceeding £40,000 were stopped. Claimants were found to have been wrongly registering as single while living with partners who had jobs or other income" (Ross 2011). It seems the *Telegraph* was (unsurprisingly) not seeking to

provide a CONSORT compliant article and its reporting does not meet CONSORT standards: for example, not adequately disclosing any pre-specified outcomes, statistical methods, summary results for each outcome, interpretation, etc.

The methodology used in this trial and the details of data analysis are not available for public scrutiny. HMRC responded to an FOI request for this by withholding this information, saying that "we believe that the requested information could be used by opportunistic individuals to make claims to which they are not entitled." However, it is hard to see how details of some fairly conventional aspects of a research project (for example, any statistical significance tests which have been conducted) could plausibly have assisted inappropriate tax credit claims. Given this nondisclosure, it appears that what Roberts (2006: 49) describes as a "bureaucratic interest in keeping secrets…deeply entrenched in bureaucratic routine" is active here.

Problematic reporting and methodology

As Gerber and Green (2012: 37) argue, if we cannot be certain that allocation procedures were robust there is a risk of selection bias; it is not clear if or how this was mitigated here. Different FOI responses from HMRC discuss a 'low risk' group, a 'medium risk' group and a 'low/medium risk' group as if the three terms refer to the same group (HMRC later stated that the 'low risk' group actually included households rated as both 'low' and 'medium' risk). HMRC did not detail how risk groups were defined or selected, and use the term 'Control Group' in the way that researchers would usually use the term 'Intervention Group'.

HMRC stated that this trial was "assessing the merits of using Credit Reference Agencies [and] we set up the control group of 750 suspicious cases to contrast the difference between the results achieved by Credit Reference Agency (CRA) data match exercises using a revised process against Business as Usual (BAU) Undeclared Partner (UP) processes." Table 1 was offered in response to an FOI request as showing "some of the key highlights".

Table 1 here

HMRC chose to select a proportion of different risk groups for the trial: stating that "samples of 150 low/medium risk and 600 high risk cases were selected from the highest award value cases…Whether the sample of high-risk cases is representative of the UK population is moot, it was representative of HMRC's focus when we scaled up the activity". It is plausible that this approach would allow predictions of the future efficacy of using credit referencing data but this depends, for example, on whether the cases in the 'control group' are likely to be similar to future cases where credit referencing data are used.

The results in Table 1 are striking: it seems that these newer processes have very substantial effects and/or the Business As Usual and Credit Reference Information groups are rather different.[xi] Without knowing about details of randomisation/selection, though, it is impossible to judge which of these is more likely. While Torgerson and Torgerson (2008: 43) argue that "[r]andomisation gives us the best method of ensuring that the variance that we do not know about or cannot measure will be balanced between the groups", one cannot tell whether randomisation was used to achieve this with this trial: the trial's results may just show differences between the groups of people in the trial.

If one bases policy on this work one is therefore left, at best, with authority-based policy: it is only due to trust in the authority of those doing the work, and HMRC's internal scrutiny processes, that this might be viewed as an adequate evidence base for policy. However, given the concerns raised above about this HMRC work, it is hard to have confidence in the quality of the aspects of this work which have not been made public and one might therefore question this authority.

Conclusions

This paper has outlined two cases where claims around data analysis and trials have served to make Government policy appear more scientific or evidence-based. This strategy was relatively successful: HMRC's analysis and use of credit referencing data was presented as a success based in part on reported trial results and, to the best of my knowledge, this paper is the first published attempt at a more critical

assessment of the quality of this work.[xii] The Behavioural Insights Team's work has also attracted considerable positive publicity.

This paper has raised questions about the lack of timely published information about some of the Behavioural Insights Team's work and about problems with and the lack of published information about an HMRC trial. Without the timely public reporting of more detail about the work which has been done, a reliance on the Behavioural Insights Team work discussed above was to a significant extent a reliance on the team's authority. The HMRC trial discussed above is, as shown by what has been released, problematic in a number of ways: the methodology and analysis are not available for public scrutiny, and the information which has been released should leave raise concerns. Clearly, the initial reporting of these trials falls far short of expectations of good practice as outlined, for example, in CONSORT guidelines.

An increased focus on evidence use in policy is welcome, and will hopefully lead to better use of relevant evidence. However, without the opportunity for more detailed analysis of and engagement with government trials, policy based on such trials should not generally be called evidence-based. In many cases, well-designed trials will provide very useful evidence. However, simply reporting that 'trials show' is of limited value without publication and critical assessment of these trials and other relevant evidence. As noted above, even with well-designed trials there is a real risk that positive findings might be a false discovery. In the cases discussed above it was not possible to conclude, based on what was published when trial findings were publicised, that these findings accurately reflect what happened there. Instead, one was – at best – relying on the authority of experts: relying on their reported findings based on a trust in their work rather than a robust assessment of this work and evidence. One might, if confident in the authority of these experts, be left at best with a relatively conventional type of authority-based policy. Policy based simply on an uncritical belief that trials show an intervention works or does not work – without adequate public assessment of or publication of these trials – should not be described as evidence-based.

Moreover, a reliance on poor quality unpublished policy trials may offer a particularly undesirable type of authority-based policy: where the trust is placed in the authority of numbers rather than the expertise of individuals. As Porter (1995) observes, it is

tempting to place trust in numbers as if the very use or presence of these numbers indicates objectivity. Daston and Galison (1992: 122) argue that quantification is part of the "ethos of restraint" and "morality" associated with objectivity. As noted above, May et al. (2005) warn of a fetishisation of randomisation in research. Rather than being a foundation for evidence-based policy, or even providing any reliable expert authority, such a trust in the mere fact that a trial has taken place could be more analogous to the way that many gamblers develop 'systems' for counting the results in games of chance such as roulette in the belief that this increases the odds of winning: a trust in numbers and the quantifiable can persist even if there is no good reason to expect these numbers to help predict future events. In such cases, trust in numbers involves sacrificing the (admittedly limited) benefits of more conventional authority-based policy, which relies on human expertise; instead, policy is based upon the authority of numbers and quantification, even when these are unreliable.

Acknowledgements

I am grateful to Stephan Köeppe for helpful comments on a draft of this paper, to Hauke Riesch for valuable suggestions re this project, and to Nicholas Fyfe and Edward Hall for useful discussions related to this work. I would also like to acknowledge the helpful comments from the journal's reviewers and editors.

I am grateful to the Behavioural Insights Team and Sense About Science for interesting discussions about this work and related topics, and to Sense About Science for helping me get in touch with the Behavioural Insights Team. This work was also assisted by the civil servants who answered my FOI requests (and related e-mails) and by My Society's invaluable whatdotheyknow.com website.

Bibliography

Advice NI, 2013, Tax Credits: Undisclosed Partner Interventions, http://www.adviceni.net/publications/PDF/Advice%20NI%20Social%20Policy%20Report%20Tax%20Credits%20Spring%202013.pdf

Amoore, L. and de Goede, M, 2008, Transactions after 9/11: the banal face of the preemptive strike, *Transactions of the Institute of British Geographers*, 33, 173-185.

Amoore, L, 2011, Data Derivatives: On the Emergence of a Security Risk Calculus for Our Times, *Theory, Culture & Society*, 28, 24-43. http://tcs.sagepub.com/content/28/6/24

Behavioural Insights Team, 2016, Academic Publications, http://www.behaviouralinsights.co.uk/academic-publications/

Cabinet Office, 2012a, Fraud, error and debt: behavioural insights team paper, https://www.gov.uk/government/publications/fraud-error-and-debt-behavioural-insights-team-paper

Cabinet Office, 2012b, Open Data White Paper: Unleashing the Potential, http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf.

Cabinet Office, 2013, OGP UK 2013 Draft National Action Plan: From Open Data to Open Government, https://www.gov.uk/government/consultations/open-government-partnership-uk-draft-national-action-plan-2013/ogp-uk-2013-draft-national-action-plan-from-open-data-to-open-government

Cartright, N, 2010, Will this Policy work for You? Predicting Effectiveness Better: How Philosophy Helps, *Templeton Foundation*, https://www.dur.ac.uk/resources/philosophy/CartwrightPSA.pdf

Cartright, N, 2013, Knowing what we are talking about: why evidence doesn't always travel, *Evidence & Policy*, 9, 1, 97-112.

Cartright, N. and Hardie, J, 2012, *Evidence-Based Policy: A Practical Guide to Doing It Better*, Oxford: OUP.

Chadborn, T, 2014, Contacting GPs to reduce unnecessary prescriptions of antibiotics, ISRCTN Registry, http://www.isrctn.com/ISRCTN32349954

Colquhoun, D, 2014, An investigation of the false discovery rate and the misinterpretation of p-values, *Royal Society Open Science*, published online ahead of print.

Daston, L. and Galison, P, 1992, The Image of Objectivity, *Representations*, 40, Special Issue: Seeing Science, 81-128.

Davies, H., Nutley, S. and Smith, P, 2000, Introducing evidence-based policy and practice in public services, in H. Davies, S. Nutley and P. Smith (eds.) *What Works?: Evidence-based Policy and Practice in Public Services*, Bristol: The Policy Press.

Gambrill, E, 1999, Evidence-Based Practice: An Alternative to Authority-Based Practice, http://www.casbrant.ca/files/upload/oacas/Chapter5/Evidence-Based_Practice.pdf

Gerber, A. G., & Green, D. P, 2012, Field experiments: Design, analysis, and interpretation. London, England: W. W. Norton.

Hall, A. and Mendel, J, 2012, Threatprints, Threads and Triggers: Imaginaries of Risk in the 'War on Terror', *Journal of Cultural Economy*, 5(1), 9-27.

Hallsworth, M., 2014, The use of field experiments to increase tax compliance, *Oxford Review of Economic Policy*, 30(4), 658-679.

Hallsworth, M., List, J. A., Metcalfe, R. D. and Vlaev, I, 2014, The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance, *National Bureau of Economic Research Working Paper*, http://www.nber.org/papers/w20007.pdf

Hallsworth, M., List, J., Metcalfe, R. and Vlaev, I., 2015, The Making of Homo Honoratus: From Omission to Commission, *National Bureau of Economic Research Working Paper*, http://www.nber.org/papers/w21210

Hallsworth, M., 2016, Personal correspondence.

Haynes, L, Service, W., Goldacre, B. and Torgerson, D, 2012, Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf.

Ioannidis, J, 2005, Why Most Published Research Findings Are False, *PLOS Medicine*, 2, 8.

Marx, J. and Hopper, F, 2005, Faith-Based versus Fact-Based Social Policy: The Case of Teenage Pregnancy Prevention, *Social Work*, 50, 3, 280-282.

May, C, Rapley, T, Moreira, T, Finch, T and Heaven, B, 2006, Technogovernance: Evidence, subjectivity, and the clinical encounter in primary care medicine, *Social Science & Medicine*, 62, 4, 1022-1030.

Roberts, A., 2006, Blacked Out: Government Secrecy in the Information Age. Cambridge, Cambridge University Press.

Ross, T, 2011, Bounty hunters called in to target benefit fraud and error, *The Telegraph*, http://www.telegraph.co.uk/news/uknews/law-and-order/8937217/Bounty-hunters-called-in-to-target-benefit-fraud-and-error.html

Schulz, K.F., Altman, D.G. and Moher, D, 2010, CONSORT 2010 statement:

updated guidelines for reporting parallel groups randomised trials. British Medical

Journal, 340: 698-702.

Sherman, L, 2013, The Rise of Evidence-Based Policing: Targeting, Testing, and Tracking, http://cebcp.org/wp-content/evidence-based-policing/Sherman-TripleT.pdf

Torgerson, D J and Torgerson, C J, 2008, Designing Randomised Trials in Health, Education and the Social Science: An Introduction. Basingstoke: Palgrave Macmillan.

Turn2Us. 2012, Tax credits: change of circumstances, compliance checks and overpayments, *Turn2Us*. http://www.turn2us.org.uk/about_us/e-bulletin/july_2012/offsetting_tax_credit_overpaym.aspx

U.S. Department of Education, 2003, Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide, http://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

What Do They Know, 2013, Behavioural Insights Team. Accessible at https://www.whatdotheyknow.com/request/behavioural_insights_team

Worral, J, 2007, Evidence in Medicine and Evidence-Based Medicine, *Philosophy Compass*, 2, 6, 981-1022.

[i] This links with the history of evidence-based medicine which, once again, has long been presented as an alternative to authority-based or eminence-based medicine.

[ii] Likewise, an evidence-based approach will be a move away from policies which – as Marx and Hopper (2005) describe – are based on religious beliefs.

[iii] Even where there is discussion of the risk of false discoveries and the benefits of replication, this may be limited. For example, the US Department of Education's (2003:8) "user-friendly guide" to evidence for educational practices appears to assume that where p=0.05 there is "only a 1 in 20 probability that the difference could have occurred by chance if the intervention's true effect is zero" and that "[r]equiring that an intervention be shown effective in two trials (or in two sites of one large trial) reduces the likelihood of such a false-positive result to 1 in 400". This guide therefore significantly underestimates the potential risk of false discoveries in some 'statistically significant' trial findings, which feeds through into a problematic understanding of replication. See also Cartright (2010: 8) for further discussion.

[iv] I was hoping to draw on these trials to consider contemporary uses of data analytics and behaviour modification.

[v] While there may be occasional cases where work cannot be made public, this should be the exception and should require special justification based on particularly compelling grounds. In the UK, a refusal to release information on research without compelling reasons conflicts with the government's argument that "Open government provides an essential foundation for economic, social and political progress, by strengthening the transparency of institutions…This transparency and accountability helps to ensure that resources are used effectively and that government, business and civil society operate in the public interest" (aCabinet Office 2013). While an 'open' government may still need to keep some information confidential for reasons of, for example, crime prevention it seems implausible that this would be the case with all the information that has been kept private in the examples discussed below.

[vi] More recently, the team has been turned into a social purpose company: owned by its employees, the UK government and the National Endowment for Science, Technology and the Arts. See http://www.behaviouralinsights.co.uk/about-us for details

[vii] Hallsworth (2014) also cites this NBER working paper in a journal article on 'The use of field experiments to increase tax compliance'. However, this journal article does not provide additional detail on how these trials were carried out. See, additionally, Hallsworth et al. (2015) for a second NBER paper based on this work.

[viii] One might also note that the Behavioural Insights Team have now moved towards pre-registering trials (see for example Chadborn 2014), while producing an impressive number of academic publications (Behavioural Insights Team 2016).

[ix] Sense About Science helpfully put me in touch with Hallsworth and Sanders at Behavioural Insights team in 2016, following my participation in Sense About Science's inquiry into non-publication of government research.

[x] Hallsworth (2016) points out that a NBER paper (Hallsworth et al. 2014) actually does offer a replication of research the Cabinet Office (2012a). This is an impressive aspect of this research. Hallsworth and Sanders (2016) also report that these trials achieved p-values which were very much lower than 0.05 and that this should indicate a lower risk of false discoveries; additionally, Hallsworth (2016) notes the very large sample size in this research. However, the limited initial publication of this research (Cabinet Office 2012a) served to make it much harder to recognise these positive features of the research. A more open and timely approach to publication would have meant the research could have been assessed more effectively – which would include evaluating its strengths.

[xi] The Business as Usual Information group generated an average yield of £228 per letter sent (142,470 letters generated a total yield of £32,489,419), while the average across the whole Control Group was £2182 per case (£1,496,164 from just 600 letters); even the 150 cases categorised as Low Risk in the Control Group yielded an average of over £200 per case.

[xii] There has, though, been some invaluable analysis of some of the harms done and concerns raised by the use of credit referencing data by Advice NI (2013) and a helpful briefing note from Turn2Us (2012).

	Business As Usual Information	Credit Reference Information Control Group Low/Medium Risk Categories[i]	Credit Reference Information Control Group High Risk Categories
Letters Issued to customers	142,470	150	600
Response Rate	34.98%	51.33%	92%
% of cases amended following customer contact	27.50%	14.3%	42.17%
*Yield [Losses prevented through intervention]	£32,489,419	£31,534	£1,496,164
**YPYC [Yield per yielding cases]	£2,371	£2,867	£5,914

Table 1

[i] HMRC originally described this as a 'Low Risk' group. However, a subsequent response indicates that this group includes both low and medium risk categories; I have therefore amended the label of this table column.